FR-TV

Full Reference Television
Phase II
Subjective Test Plan

(September 2002, Version 1.7)

CONTENTS

1 SCOPE OF THE PHASE II TEST

The main purpose of the Phase II test of VQEG is to provide input to the relevant standardisation bodies responsible for producing worldwide recommendations regarding the definition of an Objective Quality Matrix Model in the digital domain.

To perform this job, VQEG has decided to define a more precise area of interest, trying at the same time to obtain more discriminating results than those obtained in Phase I.

This gives VQEG increased motivation to pursue reliable results in a shorter period of time.

This document also defines the conditions, and the time schedule, for Phase II of FR-TV tests.

2 TEST SET-UP

This section describes the conditions and technical details according to which the VQEG FR-TV Phase II subjective tests will be carried out.

2.1 Test methodologies (DSCQS)

The selected test methodology is the Double-Stimulus Continuous Quality Scale method (DSCQS). This choice has been dictated by the fact that DSCQS is the most reliable and widely used method proposed by Rec. ITU-R BT.500-10. It should be noted that this method has been shown to have low sensitivity to contextual effects, a feature that is of particular interest considering the aim of this test.

In the DSCQS method, a subject is presented with a pair of sequences two consecutive times; one of the two sequences is the source material while the other is the test material obtained by processing the source material. The subject is asked to evaluate the picture quality of both sequences using a continuous grading scale.

The order by which the source and the processed materials are shown is random and is unknown to the subject. Subjects are invited to vote as the second presentation of the second picture begins and are asked to complete the voting in the 4 seconds after that. Usually audio or video captions announce the beginning of the sequences and the time dedicated to vote. Figure 1 shows the structure and timing of a basic DSCQS test cell.

The order of presentation of basic test cells is randomised over the test session(s) to avoid clustering of the same conditions or sequences.

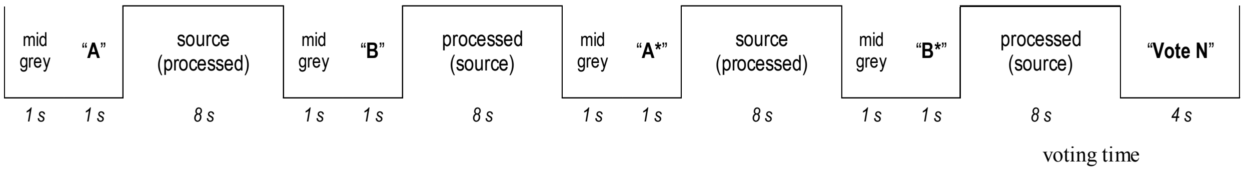


Figure 1 – DSCQS basic test cell

2.2 Grading scale

The grading scale consists of two identical 10 cm graphical scales which are divided into five equal intervals with the following adjectives from top to bottom: Excellent, Good, Fair, Poor and Bad. ITU-R Rec. 500 recognises the necessity to translate the adjectives into the language of the country where each test is performed, however it is also recognised that the use of different languages provides a slight bias due to the different meaning that each idiom gives to the translated terms. Scores will be collected on paper. The scales will be positioned in pairs to facilitate the assessment of the two sequences presented in a basic test cell. The leftmost scale will be labelled "A" and the other scale "B". To avoid loss of alignment between the votes and the basic test cells, each pair of scales will be labelled with a progressive number; in this way the subjects will have the opportunity to verify if they are expressing the current vote using the right pair of scales. The subject will be asked to record his/her assessment by drawing a short horizontal line on the grading scale at the point that corresponds to their judgement. An example of the graphical scale is provided in Annex I.

2.3 Test Material

The test will be designed using Test Video Sequences from the Video Pool (see Annex II for a definition of this and other video pools referred to in this section) that includes video material the VQEG Phase I test, video material produced by VQEG prior to the Orlando meeting, and new video material.

2.3.1 Limitations on the use of the test material

The use of the VQEG test sequences shall be restricted to the VQEG technical tests and shall not be re-used without permission for any other purpose or in any other form, including the development, promotion, demonstration and commercialisation of products directly or indirectly derived from the VQEG activities. They shall not be used without permission for any non-VQEG related evaluations, developments and/or commercial purposes (including demonstration and promotion).

Under the responsibility of the ILG, new test sequences will be used. All the entities (broadcasters, industries, research centres, normalization bodies, etc) interested in the outcome of this activity are invited to provide video test material, which they judge to be adequate to design an effective test campaign.

2.3.2. SRC Selection

The ILG will select the SRC material on the basis of the constrains listed below:

- Six (three 525 and three 625) SRCs will be selected from the VQEG Phase I Video Pool; these will be considered only in combination with three HRCs also selected from the VQEG Phase I Video Pool.
- Several SRCs will be selected from the ILG Video Pool which contains material unknown to the proponents. These SRCs will constitute about 20% of the total number of SRCs
- The rest of the SRCs will be selected from the Source Video Pool.

As a general rule the selected material will be culturally neutral and gender 'unbiased'.

2.3.3 HRC Selection

The overall selection of the HRCs will be done such that most, but not necessarily all, of the following conditions are represented:

- compression ranging from nominally 1 Mbps to 6 Mbps
- as wide a range as possible of encoders
- full and reduced horizontal resolutions

- composite and component decoder outputs
- transcoding and/or cascading processes
- image post processing, such as block reduction
- statistical multiplexing

2.3.4 HRC Verification

To be eligible for use in this test, the HRCs must meet the following technical criteria:

- maximum allowable deviation in *Peak Video Level* is +/- 10%
- maximum allowable deviation in *Black Level* is +/- 10%
- maximum allowable *Horizontal Shift* is +/- 20 pixels
- maximum allowable *Vertical Shift* is +/- 20 lines
- maximum allowable *Horizontal Cropping* is 30 pixels
- maximum allowable *Vertical Cropping* is 20 lines
- no *Vertical or Horizontal Re-scaling* is allowed
- *Temporal Alignment* between SRC and HRC sequences shall be maintained to within +/- 2 video frames.
- *Dropped or Repeated Frames* are only allowed if they do not affect temporal alignment
- no *Chroma Differential Timing* is allowed
- no *Picture Jitter* is allowed

Detailed procedures for assessing the conformance of the HRCs to the above technical criteria are given in Annex IV.

2.4 Criteria for selection of SRC x HRC combinations

A Processed Video Sequence (PVS) is a SRC processed by a HRC. For each video format (i.e., 525-line and 625-line) the test will use 64 PVSs and their corresponding source sequences, organized as follows:

- A group of 9 PVSs ordered in a 3 by 3 matrix of 3 SRCs and 3 HRCs selected from VQEG Phase I Video Pool.
- A group of 36 PVSs ordered in a 6 by 6 matrix made of 6 SRCs and 6 HRCs selected from the Video Pool, with the exclusion of material from the VQEG Phase I Video Pool.
- A group of 19 PVSs selected from the Video Pool, with the exclusion of material from the VQEG Phase I Video Pool.

Each PVS and its corresponding source sequence will have a duration of 8 seconds.

The selection of PVSs will be made, in secret, by the ILG and should conform to the following criteria:

- The expected subjective quality of PVSs should span a large range and be uniformly distributed over that range;
- Most of the selected PVSs should contain perceptible impairments.

In addition to the above criteria, the ILG will select the HRCs such that:

- PVSs drawn from the Pre-Orlando Processed Video Sequences Pool be limited to about 20% of the total number of PVSs;
- PVSs drawn from the contributions of any single proponent be limited to about 20% of the total number of PVSs;
- PVSs created by the ILG represent at least 20% of the total number of PVSs.

Note that, with some possible exceptions, the subjective tests for the 525 and 625 formats will involve different PVSs.

2.5 Distribution of tests over facilities

The tests will be distributed over three laboratories. CRC and Verizon laboratories will perform the test on 525 materials; FUB/ISCTI will perform the test on 625 materials.

In each laboratory, at least 18 subjects will participate in the test. Therefore there will be a total of 36 subjects running the 525 test and 18 subjects running the 625 tests.

2.6 Test design

Each test tape will be assigned a number so that we are able to track which facility conducts which test. The tape number will be inserted directly into the data file so that the data is directly linked to one test tape.

The test design is a full factor, balanced design to allow analysis of variance (ANOVA) consistent with the criteria outlined in Section 2.4. For each video format (i.e., 525-line, 625-line) the 64 SRC x HRC combinations will be organized into two test sessions (separated by a 15-minute break), each containing 32 basic test cells (each DSCQS basic test cell is 44 seconds long). The two test sessions will begin with 5 basic test cells containing a spread sample whose range of quality represents the quality shown in the test itself (one high quality, one low quality, three mid quality cells, presented in random order); the scores collected for these five cells will be discarded, the purpose of these being only to allow for stabilization of the viewer's responses. As a consequence of the above, each test session will be made of 37 basic test cells, resulting in a length for each test tape of 25 minutes and 40 seconds.

A viewing tape will be edited for each test session (T1 and T2). Subjects will view the 2 tapes in different ordering (T1-T2, T2-T1). Each lab should have an equal number of subjects at each ordering: 9 subjects per ordering, for a total of 18 subjects per lab.

2.7 Viewing conditions

Viewing conditions should comply with those described in Recommendation ITU-R BT.500-10. An example of a viewing room is shown in Figure 2.

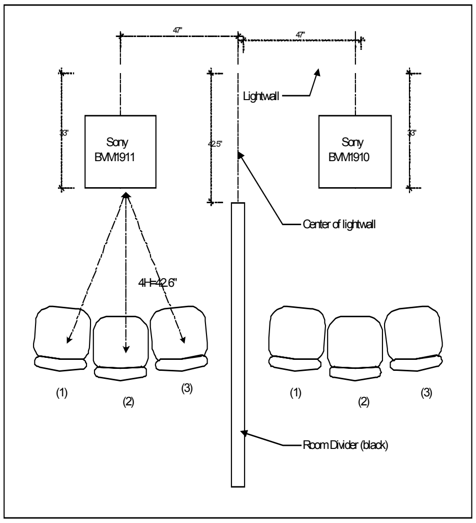


Figure 2 – Example of viewing room set-up[*]

Specific viewing conditions for subjective assessments in a laboratory environment are:

- Ratio of luminance of inactive screen to peak luminance: ≤ 0.02
- Display brightness and contrast: set up via PLUGE (see Recommendations ITU-R BT.814 and ITU-R BT.815)
- Maximum observation angle relative to the normal[**]: 30°
- Ratio of luminance of background behind picture monitor to peak luminance of picture: ≅ 0.15
- Chromaticity of background: D_{65} (0.3127, 0.3290)
- Peak screen luminance: 70 cd/m^2.
- Phosphor (x,y) chromaticities: R(0.640, 0.340), G(0.300, 0.600), B(0.150, 0.060) (these values are given in Rec. ITU-R BT.1361 and are close to both SMPTE-C and EBU values).

[*] It may become less than 0.01 when adjusted by PLUGE, but it is acceptable

[**]This number applies to CRT displays, whereas the appropriate numbers for other displays are under study.

The monitor size selected for these subjective assessments is a 19” or 20” Professional Grade monitor. In the interest of uniformity of practice and because of the availability of 19” professional-grade monitors, the 19” condition supersedes the condition specified in Rec. ITU-R BT.1129-2 for 20” and over.

The viewing distance will be 4H (i.e. four times the height of the picture tube). This choice is mainly suggested because of the strong demand for more discriminating test results.

2.8 Monitor display verification

Each subjective laboratory will undertake to ensure certain standards and will maintain records of their procedures & results, so that a flexible & usable standard of alignment can be maintained.

It is important to assure the following conditions through monitor or viewing-environment adjustment:

- To make the display conditions uniform among different facilities, no aperture correction should be used.
- Monitor bandwidth should be adequate for the displayed format
- Focus should be adjusted for maximum visibility high-spatial-frequency information
- Purity (spatial uniformity of white field) should be optimised
- Geometry should be adjusted to minimize errors & provide desired over scan. The non-active video region is defined as:
  - the top 14 frame lines
  - the bottom 14 frame lines
  - the left 14 pixels
  - the right 14 pixels

2.9 Instructions to subjects

A text (translated into the test laboratories country language) with the instructions will be read to subjects at the beginning of the first test session. An example of text to be read to the subjects before the test sessions is given in Annex I.

2.10 Viewers

One group of 18 observers will be used in each laboratory. Only non-expert viewers will participate. The term non-expert is used in the sense that the viewers' work does not involve television picture quality and they are not experienced assessors. They must not have participated in a subjective quality test over a period of six months. All viewers will be screened prior to participation for the following:

- normal (20/20) visual acuity with or without corrective glasses (per Snellen test or equivalent) [1]
- normal colour vision (per Ishihara test or equivalent)
- familiarity with the language sufficient to comprehend instruction and to provide valid responses using semantic judgement terms expressed in that language.

The results will be checked for completeness first. An observer is discarded if the number of failed votes exceeds one in one of the sessions. Additionally, the observers will be screened after the test as specified in sec. 2.3.1 of Annex 2 "Screening for DSIS, DSCQS and alternative methods except SSCQE method" of recommendation ITU-R BT.500-10. The viewers will be assigned to sub-groups, which will see the test sessions in different orders (Section 2.6). The post-test screening will NOT be applied to these sub-groups but to the groups which participate in one test as a whole. Valuable results of at least 15 viewers are required. Consequently, an additional test is necessary if the number of viewers is reduced to less than 15 as a result of the screening.

2.11 Test data collection

The collection and organization of the data files containing the votes will be under the direct responsibility of the ILG Chair.

[1] The visual acuity of a subject has to be tested for each eye separately.

3 TESTING PROCEDURES

3.1 Test schedule

TABLE 1: Below is the list of actions and the associated schedule.

Action	Done by	Source	Destination
Test plan Completed and approved	Feb. 27, 2002	VQEG	-
Final HRCs and SRCs sel.	Feb. 27, 2002	ILG	-
Call for proposal (ITU-R, ITU-T)	March 22, 2002	WP6Q SG9	Proponents
Call for expert	March 22, 2002	ILG	WP6Q SG9
Final Submission of exec. Model	November 29, 2002	Proponents	ILG
Fee payment[2]	October 18, 2002	Proponents	ILG
Video material delivery	December 10, 2002	ILG	Proponents
Tape editing	October 30, 2002	ILG	Test sites
Formal Subjective Tests	November 19, 2002	Test sites	ILG
Objective data delivery	January 31, 2003	Proponents	ILG
Statistical analysis (obj. vs subj.)	February 21, 2003	VQEG	VQEG
Final report	February 28, 2003	VQEG	WP6Q SG9

The ILG will verify that the submitted models (1) run on the ILG's computers and (2) yield the correct output values when run on the test video sequences. These verifications must be fully completed by November 29. Due to their limited resources, the ILG may encounter difficulties verifying executables submitted too close to the November 29 model submission deadline. Therefore, proponents are *strongly* encouraged to submit a prototype model to the ILG before October 18, to work out platform compatibility problems well ahead of the drop-dead date. Proponents are also *strongly* encouraged to submit their final model executable by November 15, giving the ILG two weeks to resolve problems arising from the verification procedure.

The ILG requests that proponents kindly estimate the run-speed of their executables on a test video sequence and to provide this information to the ILG.

3.2 Results analysis

The results will be analysed, in conjunction with the objective data provided by the proponents, to derive the evaluation metrics of Section 5. These metrics are calculated by each proponent and verified by the ILG, or they may be calculated completely by the ILG and verified by the proponents. The results will be reported anonymously to the outside world but identified by proponent to VQEG. Peak

[2] Payment will be made directly from each proponent to the selected testing facility, according to a table agreed by ILG and distributed to the proponents.

signal to noise ratio (PSNR) shall be reported if calculated by a VQEG member. It will be based on normalization of gain, offset and spatial alignment implemented by an agreed method and to an agreed accuracy.

4 OBJECTIVE QUALITY MODEL SUBMISSION REQUIREMENTS

The objective quality model submissions must conform to the submission requirements as specified by the VQEG Phase I full reference objective test plan (e.g., original and processed video file format, model results file format). However, unlike the Phase I test, there will be no normalization performed on the processed video sequences (i.e., no correction for gain and level offset, spatial shifts, or temporal shifts). If normalization is required by the submitted model, this must be performed by the binary executable submitted to VQEG. Accordingly, test video sequences will contain no information relative to normalization. Specifically, the first and last second of the video sequence files will not contain an alignment pattern to facilitate the normalization operation. Thus, a complete sequence on the computer will be:

VideoNotUsed (10 frames) + Video (8 sec)_+ VideoNotUsed (10 frames)

4.1 Video data format, general

Objective models will take two Rec. 601 digital video sequences as input, here referred to as Source and Processed, with the goal of predicting the quality difference between the Source and Processed sequences. The video sequences will be in either 625/50 or 525/60 format.

4.2 Video data format, specifics

The test video sequences will be in ITU Recommendation 601 4:2:2 component video format using an aspect ratio of 4:3. This may be in either 525/60 or 625/50 line formats. The temporal ordering of fields F1 and F2 will be described below with the field containing line 1 of (stored) video referred to as the Top-Field.

Data storage:
A LINE: of video consists of 1440 8 bit data fields in the multiplexed order: Cb Y Cr [Y]... . Hence there are 720 Y's and 360 Cb's and 360 Cr's per line of video.

A FRAME: of video consists of 486 active lines for 525/60 Hz material and 576 active lines for 625/50 Hz material. Each frame consists of two interlaced Fields, F1 and F2. The temporal ordering of F1 and F2 can be easily confused due to cropping and so we make it specific as follows:
For 525/60 material: F1--the Top-Field-- (containing line 1 of FILE storage) is temporally LATER (than field F2). F1 and F2 are stored interlaced.
For 625/50 material: F1--the Top-Field-- is temporally EARLIER than F2.
The Frame SIZE:
for 525/60 is: 699840 bytes/frame,
for 625/50 is: 829440 bytes/frame.

A FILE: is a contiguous byte stream composed of a sequences of frames.
Due to the choice of making the test on sequences each of them 8 seconds long, the files will thus have a total byte count of
for 525/60: 240 frames x 699.840 bytes/frame = 165.888.000 bytes/sequence,
for 625/50: 200 frames x 829.440 bytes/frame = 167.961.600bytes/sequence

Multiplex structure: Cb Y Cr [Y] ... 1440 bytes/line
720 Y's/line
360 Cb's/line
360 Cr's/line

TABLE 2: Format summary

	-- 525/60 --	-- 625/50 --
active lines	486	576
frame size (bytes)	699.840	829.440
fields/sec (Hz)	60	50
Top-Field (F1)	LATER	EARLIER
8seconds Seq-length (bytes)	167.961.600	165.888.000
8 seconds + 20 frames Seq-length (bytes)	181.958.400	182.476.800

4.3 Model input and output data format

For each video format, 525/60 and 625/50, the model will be given a ASCII file listing pairs of video sequence files to be processed. Each line of this file has the following format:

<source-file> <processed-file>

where <source-file> is the name of a source video sequence file and <processed-file> is the name of a processed video sequence file, whose format is specified in section 4.2 of this document. File names may include a path. For example, an input file for the 525/60 cases might contain the following:

```
/video/V2src1_525.yuv    /video/V2src1_hrc2_525.yuv
/video/V2src1_525.yuv    /video/V2src1_hrc1_525.yuv
/video/V2src2_525.yuv    /video/V2src2_hrc1_525.yuv
/video/V2src2_525.yuv    /video/V2src2_hrc2_525.yuv
```

The output file is an ASCII file created by the model program, listing the name of each processed sequence and the resulting Video Quality Rating (VQR) of the model. The contents of the output file should be flushed after each sequence is processed, to allow the testing laboratories the option of halting a processing run at any time. Each line of the ASCII output file has the following format:

<processed-file> VQR4H

Where <processed-file> is the name of the processed sequence run through this model, without any path information. VQR4H is the Video Quality Ratings produced by the objective model for the 4H viewing distances. For the input file example, this file contains the following:

```
V2src1_hrc2_525.yuv  0.150
V2src1_hrc1_525.yuv  1.304
V2src2_hrc1_525.yuv  0.102
```

V2src2_hrc2_525.yuv 2.989

Each proponent is also allowed to output a file containing Model Output Values (MOVs) that the proponents consider to be important. The format of this file will be

V2src1_hrc2_525.yuv 0.150 MOV$_1$ MOV$_2$,... MOV$_N$

V2src1_hrc1_525.yuv 1.304 MOV$_1$ MOV$_2$,... MOV$_N$

V2src2_hrc1_525.yuv 0.102 MOV$_1$ MOV$_2$,... MOV$_N$

V2src2_hrc2_525.yuv 2.989 MOV$_1$ MOV$_2$,... MOV$_N$

All video sequences will be displayed in over-scan and the non-active video region is defined in Section 2.8.

4.4 Final submission of executable model

For each video format, 525/60 and 625/50, a set of 2 source and processed video sequence pairs will be used as test vectors. They will be available for downloading on the VQEG web site http://www.vqeg.org/.

Each proponent will send an executable of the model and the test vector outputs to the CRC and FUB/ISCTI laboratories by the date specified in action item "Final submission of executable model" of Section 3.1. The executable version of the model must run correctly on one of the two following computing environments:

- SUN SPARC workstation running the Solaris 2.3 UNIX operating system (SUN OS 5.5).
- WINDOWS 2000 workstation.

The use of other platforms will have to be agreed upon with the independent laboratories prior to the submission of the model.

The independent laboratories will verify that the software produces the same results as the proponent with a maximum error of 0.1%. If greater errors are found, the independent and proponent laboratories will work together to correct them. If the errors cannot be corrected, then the ILG will review the results and recommend further action.

4.5 Test sequence objective analysis

Each proponent will receive Source and Processed video sequences to be used in the test by the date specified in action item "Video material delivery" of Section 3.1. Each proponent will send the objective data to the ILG by the date specified in action item "Objective data delivery" of Section 3.1.

The independent laboratories will verify the objective data provided by each proponent. Specifically, the independent laboratories will verify that each model produces the same results as those submitted by the proponent, within an acceptable error of 0.1%. This verification will be limited to a randomly selected subset (about 10%) of source and processed video sequence pairs. The random sequence subset will be selected by the ILG and kept confidential. If errors greater than 0.1% are found, then the independent and proponent laboratories will work together to discover the source of the errors. If processing and handling errors are ruled out, then the ILG will recommend further action.

The outputs by the objective video quality model (the VQR's) should be correlated with the viewer Difference Mean Opinion Scores (DMOS's) in a predictable and repeatable fashion. It is expected that the VQR's and DMOS's will be positive in typical situations and increasing values will predict increasingly perceptible differences between Source and Processed sequences. Negative values of both may occur in certain situations and will be allowed.

5 OBJECTIVE QUALITY MODEL EVALUATION CRITERIA

5.1 Introduction to evaluation metrics

A number of attributes characterize the performance of an objective video quality model as an estimator of video picture quality in a variety of applications. These attributes are listed in the following sections as:

- Prediction Accuracy
- Prediction Monotonicity
- Prediction Consistency

This section lists a set of metrics to measure these attributes. The metrics are derived from the objective model outputs and the results from viewer subjective rating of the test sequences. Both objective and subjective tests will provide a single number (figure of merit) for each Source and Processed sequence pair that correlates with the video quality difference between the Source and Processed sequences. It is presumed that the subjective results include mean ratings and error confidence intervals that take into account differences within the viewer population and differences between multiple subjective testing laboratories.

5.2 Prediction nonlinearity

The outputs by the objective video quality model (the VQR's) should be correlated with the viewer Difference Mean Opinion Scores (DMOS's) in a predictable and repeatable fashion. The relationship between predicted VQR and DMOS need not be linear as subjective testing can have nonlinear quality rating compression at the extremes of the test range. It is not the linearity of the relationship that is critical, but the stability of the relationship and a data set's error-variance from the relationship that determine predictive usefulness. To remove any nonlinearity due to the subjective rating process (see Figure 3) and to facilitate comparison of the models in a common analysis space, the relationship between each model's predictions and the subjective ratings will be estimated using a nonlinear regression between the model's set of VQR's and the corresponding DMOS's.

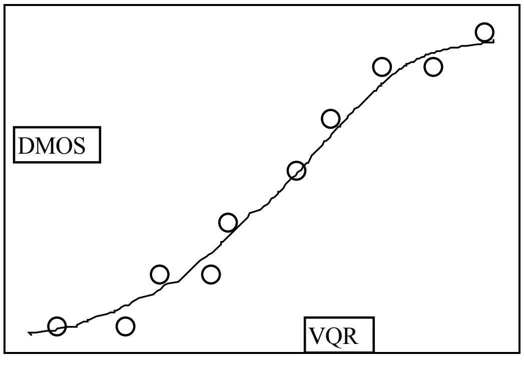


Figure 3. Example Relationship between VQR and DMOS

The nonlinear regression will be fitted to the [VQR,DMOS] data set and be restricted to be monotonic over the range of VQR's. The functional form of the nonlinear regression is not critical except that it be monotonic, reasonably general, and have a minimum number of free parameters to avoid over-fitting of the data.

The functional form to be regressed is the 5-parameter logistic curve:

$DMOS_p(VQR) = A0 + (A1-A0)/(1 + ((X+A4)/A2)\char`^A3)$

fitted to the data [VQR,DMOS].

The nonlinear regression function will be used to transform the set of VQR values to a set of predicted MOS values, $DMOS_p(VQR)$, which will then be compared with the actual DMOS values from the subjective tests.

Besides carrying out an analysis on the mean one can do the same analysis on the individual Opinion Scores (OS), leading to individual Differential Opinion Scores (DOS). This has the advantage of taking into account the variations between subjects. For objective models there is no variance and thus $OS_p= MOS_p$ and $DOS_p=DMOS_p$.

5.3 Evaluation metrics

This section lists the evaluation metrics to be calculated on the subjective and objective data. Once the nonlinear transformation of section 5.2 has been applied, the objective model's prediction performance is then evaluated by computing various metrics on the actual sets of subjectively measured DMOS and the predicted $DMOS_p$. The set of differences between measured and predicted DMOS is defined as the quality-error set Qerror[]:

$Qerror[i] = DMOS[i] – DMOS_p[i]$

where the index 'I' refers to an individual processed video sequence.

Metrics relating to Prediction Accuracy of a model

Metric 1: The Pearson linear correlation coefficient between $DMOS_p$ and DMOS.

Metrics relating to Prediction Monotonicity of a model

Metric 2: Spearman rank order correlation coefficient between $DMOS_p$ and DMOS.

VQR performance can also be assessed by correlating subjective scores and corresponding VQR predicted scores after the subjective data have been averaged over both subjects and SRC's. Such an analysis has been proposed and is under consideration by VQEG.

Metrics relating to Prediction Consistency of a model

Metric 3: Outlier Ratio of "outlier-points" to total points N.

Outlier Ratio = (total number of outliers)/N

where an outlier is a point for which: ABS[Qerror[i]] > 2*DMOSStandardError[i].

Twice the DMOS Standard Error is used as the threshold for defining an outlier point.

Metric 4, 5, 6: Evaluation based on T1.TR.PP.72-2001 ("Methodological Framework for Specifying Accuracy and Cross-Calibration of Video Quality Metrics")

4. RMS Error,

5. Resolving Power, and

6. Classification Errors

Evaluation of models using the T1A1 method will omit the cross-calibration procedure described therein, as it is not relevant to measures of performance of individual models.

5.4 Generalizability

Generalizability is the ability of a model to perform reliably over a very broad set of video content. This is obviously a critical selection factor given the very wide variety of content found in real applications. There is no specific metric that is specific to generalizability so this objective testing procedure requires the selection of as broad a set of representative test sequences as is possible. The test sequences and specific HRC's will be selected by the experts of the VQEG's Independent Laboratory Group and should ensure broad coverage of typical content (spatial detail, motion complexity, color, etc.) and typical video processing conditions. The breadth of the test set will determine how well the generalizability of the models is tested. At least 12 different scenes are recommended as a minimum set of test sequences.

5.5 Complexity

The performance of a model as measured by the above Metrics #1-6 will be used as the primary basis for model analysis. The specification of model complexity, while potentially important, is not in the scope of this test. This information can be requested from the proponents.

6 CONCLUSIONS

All results will be reported in a technical document and submitted to the Standardization bodies.

Discussion will be provided to help with the interpretation of the results.

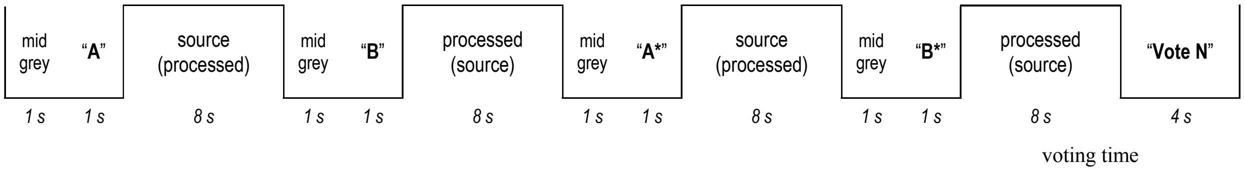
Annex I
Instructions to the Subjects

"In this test, we ask you to evaluate the overall quality of the video material you see. We are interested in your opinion of the video quality of each scene. Please do not base your opinion on the content of the scene or the quality of the acting. Take into account the different aspects of the video quality and form your opinion based upon your total impression of the video quality.

Possible problems in quality include:

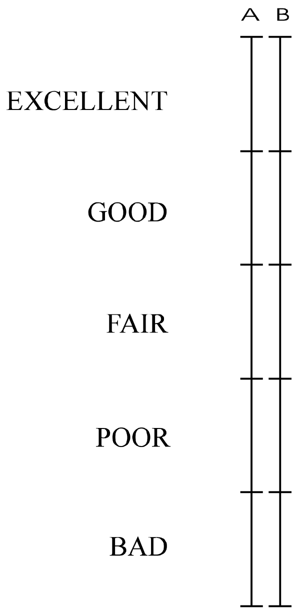
- poor, or inconsistent, reproduction of detail;
- poor reproduction of colours, brightness, or depth;
- poor reproduction of motion;
- imperfections, such as false patterns, or "snow".

The test consists of a series of judgement trials. During each trial, two versions of a single video sequence, which may or may not differ in picture quality, will be shown in the following way:



"A" is the first version, "B" is the second version. The first presentation of a trial will be announced with the written caption "A", and the second with "B". This pair of presentations will then be repeated, thereby completing a single trial.

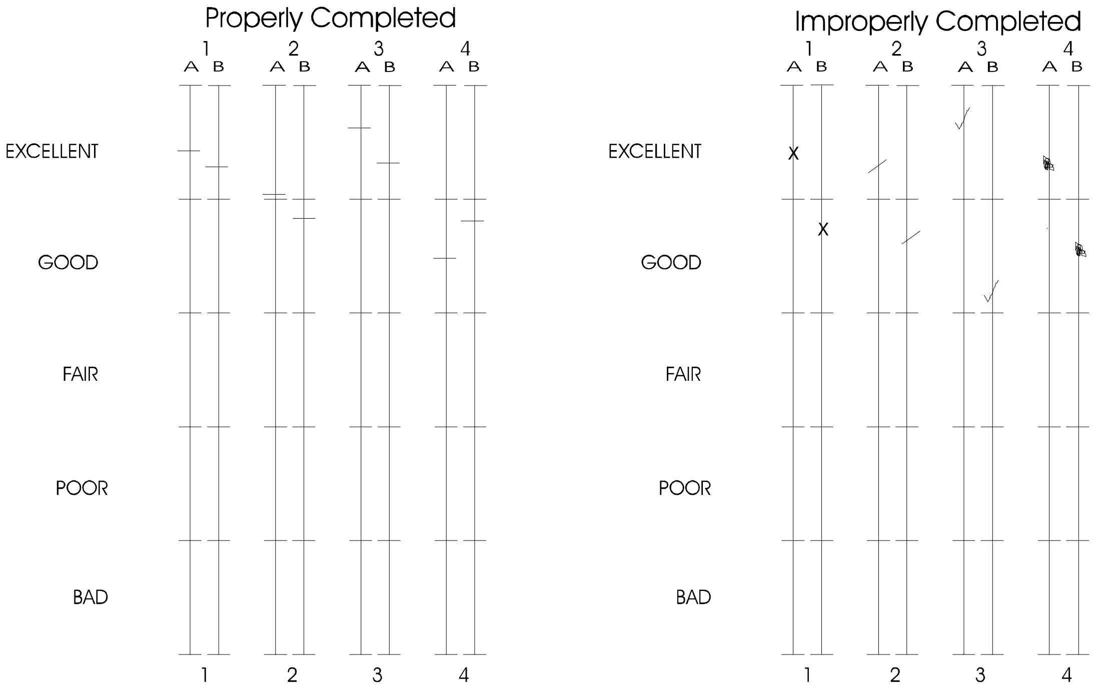
In judging the overall quality of the presentations, we ask you to use judgement scales like the samples shown below.



SAMPLE QUALITY SCALE

As you can see, there are two scales for each trial, one for the “A” presentation and one for the “B” presentation, since both the “A” and “B” presentations are to be judged.

The judgement scales are continuous vertical lines that are divided into five segments. As a guide, the adjectives “excellent”, “good”, “fair”, “poor”, and “bad” have been aligned with the five segments of the scales. You are asked to place a single horizontal line at the point on the scale that best corresponds to your judgement of the overall quality of the presentation (as shown in the example).



You may make your mark at any point on the scale, which most precisely represents your judgement.

In making your judgements, we ask you to use the first pair of presentations in the trial to form an impression of the quality of each presentation, but to refrain from recording your judgements. You may then use the second pair of presentations to confirm your first impressions and to record your judgements in your Response Booklet.

We will now show you four demonstration trials.

DEMONSTRATION TRIALS PRESENTED AT THIS POINT

Annex II
Video Material Definitions

Video Pool

Video Pool is a set of video sequences, from which **Test Video Sequences** will be selected by the ILG group. The Video Pool consists of video sequences drawn from the **Source Video Pool,** the **Pre-Orlando Processed Video Sequences Pool,** the **Proponent Processed Video Sequences Pool**, the **VQEG Phase I Video Pool**, and the **ILG Video Pool.**

Test Video Sequences

Test Video Sequences are a set of video sequences that are selected secretly by the ILG group and used to evaluate objective models submitted by proponents. The source of the Test Video Sequences will be disclosed when they are delivered to each proponent.

Source Video Pool

The Source Video Pool contains source sequences only. The Source Video Pool is available to any proponent who requests and pays appropriate expenses for a copy of this material. The entire content of this Source Video Pool is listed in the two following tables, for 525/60 and 625/50 formats. Note that the video material differs for the two formats. Accordingly, the subjective tests for the 525/60 and 625/50 formats will involve different video materials.

Source Video Pool for 525/60			
Scene	**Description**	**Origination**	**Duration** (hh:mm:ss)
“Apollo 13”	Lift off scene: synthetic picture, fine detail, jerky motion	Original Film, telecined to 480i60 Universal Studios; POC: Teranex	00:01:40
“Casper”	Synthetic picture-digital CGI	12 fps original converted to film at 24 fps, telecined to 480i60 Universal Studios; POC: Teranex	00:01:05
“Woody Woodpecker”	Synthetic picture-traditional animation	12 fps original converted to film at 24 fps, telecined to 480i60 Universal Studios; POC: Teranex	00:01:10
“Land Before Time”	Synthetic picture	Original Film, telecined to 480i60 Universal Studios, POC: Teranex	00:01:10
“The Thing”	Remake of original, Snow scenes, various Motion	Original Film, telecined to 480i60 Universal Studios, POC: Teranex	00:02:40
“Frankenstein"”	Black and white original, “Bringing to life” scene	Original Film, telecined to 480i60 Universal Studios, POC: Teranex	00:02:10

“Mummy Returns”	Movie Trailer-special effects	Original Film, telecined to 480i60 Universal Studios, POC: Teranex	00:01:40
Ballet Dancing	Indoor Ballet Dancing Couple, fast rapid movement	Original Film, telecined to 480i60 Kodak; POC: Teranex	00:01:50
Universal Theme Park	Varying motion, high contrast, full sunlight, water rides, inside rides, roller coaster	Capture with DigiBetaCam Teranex; POC: Teranex	00:08:25
VQEG Phase 1 scenes			00:01:20
“Sahara”	Natural scenery, bugs, reptiles, sand storm, waterfall, nocturnal animals, fine detail	Original Film/HiDef—HD Down (3/2) insertion Mandalay Media Arts; POC: Teranex	00:26:55
Total Time			**00:50:05**

Source Video Pool for 625 / 50

Scene	**Description**	**Origination**	**Duration** (hh:mm:ss)
Sahara 1	Large surfaces crossed by slowly moving objects.	Original Film/HiDef—HD Down (3/2) insertion Mandalay Media Arts; POC: Teranex	00:03:00
Universal studio 1	Slow moving objects on detailed background	Capture with DigiBetaCam Teranex; POC: Teranex	00:03:00
Universal studio 2	Cartoons, objects with curved drawings	Capture with DigiBetaCam Teranex; POC: Teranex	00:03:00
Football with ladies	Quick moving objects in external lit area	Captured in D5 German Broadcaster SWR/ARD; POC Teranex	00:03:00
Sahara 2	Moving objects with details in half sun light exposure	Original Film/HiDef—HD Down (3/2) insertion Mandalay Media Arts; POC: Teranex	00:03:00
New York	Slow moving geometric objects with details	Original Film (16:9), telecined to 576i50 German Broadcaster; POC Teranex	00:03:00
Universal studio 3	Fixed and moving objects of various colors	Capture with DigiBetaCam Teranex; POC: Teranex	00:03:00
Volleyball	Quick moving objects with interior lighting	Captured in D5 German Broadcaster SWR/ARD; POC Teranex	00:03:00
Dancing	Dancers on square wooden floor	Captured in D5 German Broadcaster SWR/ARD; POC Teranex	00:03:00
Total Time			**00:27:00**

Processed Video Sequences Pool Provided By Proponents

A Processed Video Sequence (PVS) is a SRC processed by a HRC. The following table consists of all PVSs provided by the proponents to the ILG (the ILG has produced other secret PVSs that are not included in this table).

Table of Pre-Orlando Processed Video Sequences						
Bit Rate (Mbps)	**Modulation**	**Distortion**	**Output**	**Proponent**	**525**	**625**
6.0 @ 720H	64QAM	23.5 dB noise	601	R&S	x	
4.0 @ 720H	64QAM	QEF	601	TDF		x
3.0 @ 720H	64QAM	QEF	601	R&S	x	
3.0 @ 320H		QEF	601	BT		X
3.0 @ 320H		QEF	601	BT	X	
3.0 @ 704H	64QAM	21.6dB noise	601	R&S	X	
3.0 @ 704H	64QAM	QEF	601	TDF		X
3.0 @ 704H	64QAM	QEF	PAL	TDF		X
3.0 @ 528H	64QAM	QEF	601	TDF		X
2.5 @ 704H	64QAM	QEF	601	R&S	X	
2.0 @ 704H	64QAM	QEF	601	TDF		x
2.0 @ 720H	64QAM	QEF	601	R&S	x	
2.0 @ 720H	64QAM	QEF	NTSC	NTIA	x	
2.0 @ 528H	64QAM	QEF	NTSC	NTIA	x	
2.0 @ 720H, cascaded, 4→2		QEF	601	BT	X	
4→2 transcoded @ 704H	64QAM	QEF	601	TDF		x
1.5 @ 720H	64QAM	QEF	601	R&S	X	
1.5 @ 704H	64QAM	QEF	601	R&S	X	
1.5 @ 528H	64QAM	QEF	NTSC	NTIA	X	
1.0 @ 704H	64QAM	QEF	601	R&S	X	
1.0 @ 320H		QEF	601	BT		X
1.0 @ 320H		QEF	601	BT	X	
1.0 @ 320H, cascaded, 3→1		QEF	601	BT		X

1.0 @ 320H, cascaded, 3→1		QEF	601	BT	X	
1.0 @ 528H	64QAM	QEF	NTSC	NTIA	X	
1.0 @ 352H	64QAM	QEF	NTSC	NTIA	X	

Proponent Processed Video Sequence Pool

The Proponent Processed Video Sequence Pool consists of new PVSs generated by a proponent after the VQEG Orlando meetings of July 2001. In generating these proponent PVS, the **Source Video Pool**, in its entirety, for the given video format, must be used as input to the HRC.

VQEG Phase I Video Pool

The VQEG Phase I Video Pool consists of SRC and PVS used in Phase I of the VQEG project.

ILG Video Pool

The ILG Group Video Pool consists of new SRCs and new PVSs created by the ILG group. These will remain unknown to the proponents.

Annex III
Objective Video Quality Model Attributes

This test plan presents several important attributes, and supporting metrics, that relate to an objective quality model's ability to predict a viewer's rating of the difference between two video sequences. This annex provides further background on the nature of these attributes, and serves as a guide to the selection of metrics appropriate for measuring each attribute. The discussion is in terms of the relation between the subjective DMOS data and the model's transformed $\mathrm{DMOS_p}$ data. The schematic data and lines are not real, but idealized examples only meant to illustrate the discussion. In the interest of clarity, only a few points are used to illustrate the relationship between objective $\mathrm{DMOS_p}$ and subjective DMOS, and error bars on the subjective DMOS data are left out.

Attribute1: Prediction Accuracy

This attribute is simply the ability of the model to predict the viewers' DMOS ratings with a minimum error "on average". The model in Figure 4 is seen to have a lower average error between $\mathrm{DMOS_p}$ and DMOS than the model in Figure 5, and has therefore greater prediction accuracy.

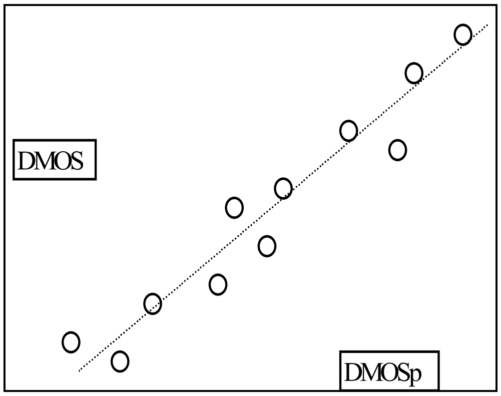


Figure 4. Model with greater accuracy

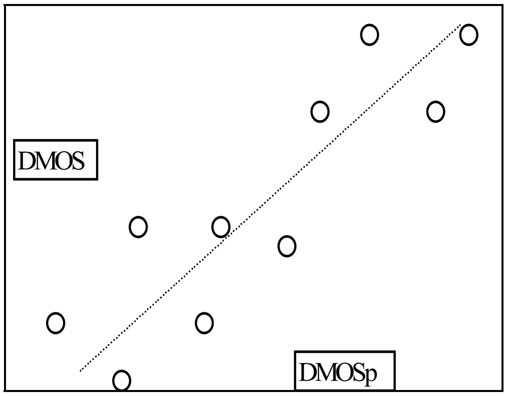


Figure 5. Model with lower accuracy

A number of metrics can be used to measure the average error, with root-mean-square (RMS) error being a common one. In order to incorporate the known variance in subjective DMOS data, the simple RMS error can also be weighted by the confidence intervals for the mean DMOS data points. The Pearson linear correlation coefficient, although not a direct measure of average error magnitude, is another common metric that is related to the average error in that lower average errors lead to higher values of the correlation coefficient.

Attribute2: Prediction Monotonicity

An objective model's $\mathrm{DMOS_p}$ values should ideally be completely monotonic in their relationship to the matching DMOS values. The model should predict a change in $\mathrm{DMOS_p}$ that has the same sign as the change in DMOS. Figures 6 and 7 below illustrate hypothetical relationships between $\mathrm{DMOS_p}$ and DMOS for two models of varying monotonicity. Both relationships have approximately the same prediction accuracy in terms of RMS error, but the model of Figure 6 has predictions that monotonically increase. The model in Figure 7 is less monotonic and falsely predicts a decrease in $\mathrm{DMOS_p}$ for a case in which viewers actually see an increase in DMOS.

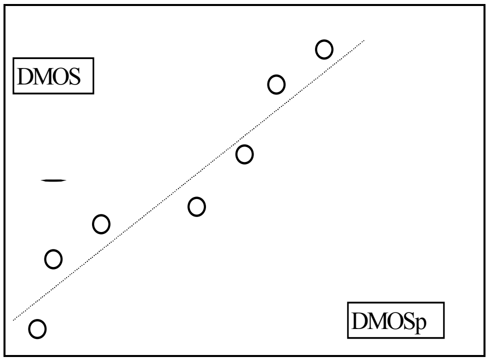


Figure 6. Model with more Monotonicity

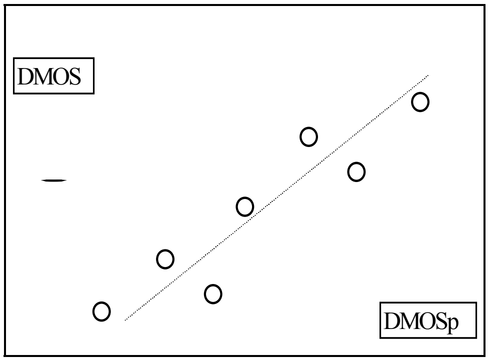


Figure 7. Model with less Monotonicity

The Spearman rank-order correlation between $DMOS_p$ and DMOS is a sensitive measure of Monotonicity. It also has the added benefit that it is a nonparametric test that makes no assumptions about the form of the relationship (linear, polynomial, etc.). Another method to understand model Monotonicity is to perform pair-wise comparisons on HRC's by type of sequence, bit-rate, and any other parameters defining an HRC). The change between the pairs in DMOS should correlate with the change in $DMOS_p$.

Attribute3: Prediction Consistency

This attribute relates to the objective quality model's ability to provide consistently accurate predictions for all types of video sequences and not fail excessively for a subset of sequences.

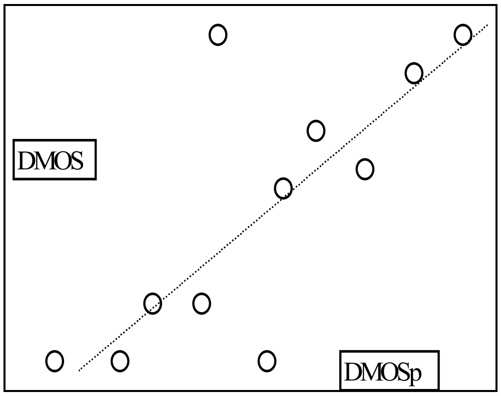


Figure 8. Model with large outlying errors

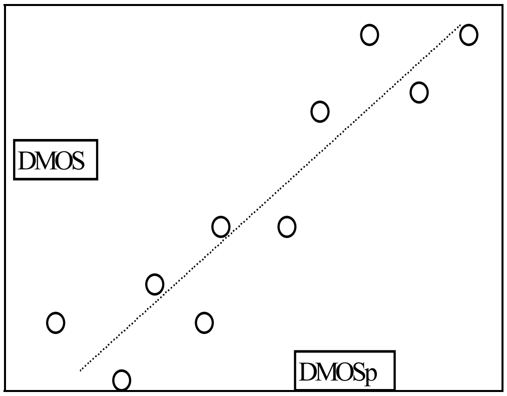


Figure 9. Model with consistent errors

Figures 8 and 9 show models with approximately equal RMS errors between predicted and measured DMOS. Figure 8 is an example of a model that has quite accurate predictions for the majority of sequences but has large prediction error for the two points in the middle of the figure. Figure 9 is an example of a model that has a balanced set of prediction errors - it is not as accurate as the model of Figure 8 for most of the sequences but it performs "consistently" by providing reasonable predictions for all the sequences. The model's prediction consistency can be measured by the number of outlier points (defined as having an error greater than a given threshold such as one confidence interval) as a fraction of the total number of points. A smaller outlier fraction means the model's predictions are more consistent.

Annex IV
Conformance of HRC to Technical Criteria

These measurements make use of the Tektronix WFM601M waveform monitor, or equivalent. They use the 75% colorbar leader that appears on the SRC and HRC tapes. Also, freeze frames from the sequence Apollo (4:3 aspect) and Return of the Mummy (letterboxed) are required. Good freeze frames are ones with bright areas along the raster edges.

A freeze frame output from the D1 VTR can be obtained by accessing the SYSTEM menu. From there, check the Standby Off column in the Statemap, and set the Video parameter to PB. Ensure as well that the D1 freeze mode is set to Frame. This parameter can be set in the VIDEO menu.

1: Obtain reference readings from the SRC tape

1-1: Measure Peak Video Levels

With the WFM601 in Waveform mode, set for RGB display, with the Low Pass filter active, measure the amplitude of the peak white bar of the SRC colorbar leader, and obtain values for RGB.

1-2: Measure Black Level

With the WFM601 in Waveform mode, set for RGB display, with the Low Pass filter active, measure the amplitude of the black bar or the SRC colorbar leader for the G channel.

1-3: Measure Colorbar Transitions for Luminance

Set the WFM601M to DGTL WFM mode, and display only the Y channel. Using the sample cursor, determine the sample numbers for the colorbar transitions at 2 places: white to yellow, and red to blue. These transitions generally occur over a 4-step interval, so use the fourth step as the transitional sample number.

1-4: Measure Colorbar Transitions for Chrominance

Set the WFM601M to DGTL WFM mode, and display only the Cr channel. Using the sample cursor, determine the sample numbers for the colorbar transitions at 2 places: white to yellow, and red to blue. These transitions generally occur over a 4 step interval, so use the forth step as the transitional sample number.

Repeat this procedure for the Cb channel.

1-5: Determine First and Last Lines for Letterboxed Sequence

Use a freeze frame from the sequence Return of the Mummy. Set the WFM601M to DGTL WFM mode, and display only the luminance channel. Select an appropriate sample value, with bright areas at top and bottom of raster, and use the Line Select feature to determine the line numbers for the start and end of active lines in field 1.

1-6: Ensure Absence of Jitter

Using the same sequence as section 1-5, ensure that the values for first and last lines do not change over time (a few seconds).

1-7: Determine First and Last Lines for Non-Letterboxed Sequence

Use a suitable freeze frame from the sequence Apollo. Set the WFM601M to DGTL WFM mode, and display only the luminance channel. Select an appropriate sample value, with bright areas at top and bottom, and use the Line Select feature to determine the line numbers for the start and end of active lines in field 1. Determine the total number of lines for field 1 that contain active video.

1-8: Determine the Number of Active Horizontal Samples

Use the same freeze frame from the sequence Apollo. Set the WFM601M to DGTL WFM mode, and display only the luminance channel. Select an appropriate line, with bright areas at each edge, and use the Sample Select feature to determine the first and last sample numbers that contain active video. Determine the total number of Y samples in the line that contain active video.

2: Obtain readings from each HRC tape

For each HRC tape, perform the measurements from section 1, using the same (or close) freeze frames.

2-1: Measure Peak Video Levels

2-2: Measure Black Level

2-3: Measure Colorbar Transitions for Luminance

2-4: Measure Colorbar Transitions for Chrominance

2-5: Determine First and Last Lines for Letterboxed Sequence

2-6: Ensure Absence of Jitter

2-7: Determine First and Last Lines for Non-Letterboxed Sequence

2-8: Determine the Number of Active Horizontal Samples

The measurements are summarized in the following table:

Test Signal	Item	SRC	HRC
White from 75% colorbar leader	Peak Video Level (RED)(mv)		
	Peak Video Level (GREEN)(mv)		
	Peak Video Level (BLUE)(mv)		
Black from 75% colorbar leader	Black Level (GREEN)(mv)		
Colorbar White to Yellow transition	Y transition sample #		
	Cr transition sample #		
	Cb transition sample #		
Colorbar Red to Blue transition	Y transition sample #		
	Cr transition sample #		
	Cb transition sample #		
Letterboxed Freeze Field 1	First active line #		
	Last active line #		
Letterboxed additional freezes	Check for Jitter		
Full Screen Freeze, Field 1	First active line #		
	Last active line #		
Full Screen Freeze, Field 2	First active line #		
	Last active line #		
Calculated Value	Total active Lines		
Full Screen Freeze, Field 1	First active sample #		
	Last active sample #		
Calculated Value	Total active samples		

3: Assess and record the results as follows:

3-1: Check Peak Video Levels

Compare the SRC and HRC levels obtained for RGB peak values. Report the difference value as a percentage of 700.

3-2: Check Black Levels

For HRCs with component 601 outputs, Black Level should equal 0mv. Report any deviation as a percentage of 700.

3-3: Check for Horizontal Re-scaling and Horizontal Shift
Compare the SRC and HRC Colorbar Transition sample values obtained in step 1-3, and step 2-3. These should either be equal, or offset by an equal amount. (*Allow a measurement tolerance of 1 Y sample*). If the offset value is not constant, then horizontal re-scaling has occurred, and the HRC should be rejected. If the value is constant, then report that value in Y samples, as the Horizontal Shift.

3-4: Check for Chroma/Luma Differential Timing
Compare the SRC and HRC Colorbar Transition sample values obtained in step 1-4, and step 2-4. These should either be equal, or offset by an equal amount. (*Allow a measurement tolerance of 1 Cr / Cb sample*). If the offset value is not constant, and if no re-scaling has been detected in step 3-3, then Differential Timing is present, and the HRC should be rejected.

3-5: Check for Vertical Re-scaling and Vertical Shift
Compare the SRC and HRC start and end of field 1 for the letterboxed sequence. These should either be equal, or offset by an equal amount. (*Allow a measurement tolerance of 1 line*). If the offset value is not constant, then vertical rescaling has occurred, and the HRC should be rejected. If the value is constant, then report that value as the Vertical Shift.

3-6: Check for Jitter
Make this determination (yes/no) from steps 1-6 and 2-6.

3-7: Check for Vertical Cropping
Compare the number of active video lines in field 1 of the SRC and HRC, as obtained in steps 1-7 and 2-7. Determine the SRC minus HRC total cropped lines value for field 1 and double this to account for field 2.

3-8: Check for Horizontal Cropping
Compare the number of horizontal active samples for the SRC and HRC sequences, obtained in steps 1-8 and 2-8. Report the SRC minus HRC Total samples value, in Y samples.

3-9: Check for Frame Alignment due to Dropped or Repeated Frames
The check for dropped frames should be done during the editing process, as it is unpredictable and could occur in some sequences and not others. Frame repeats are allowed as long as they do not affect temporal alignment (i.e. the segment of video after the frame repeat has the same temporal alignment as the segment of video before the frame repeat).

A reporting table could look like the following:

	Reporting value	Tolerance	HRC Tape 1
Video Level Gain (red)	+/- %	+/- 70mv	
Video Level Gain (green)		+/- 70mv	
Video Level Gain (blue)		+/- 70mv	
Black Level Gain		+/- 70mv	
Horizontal Re-Scaling	yes / no	must equal NO	
Horizontal Shift	+/- # of Y samples	20 pixels	
Chroma Diff Timing	yes / no	must equal NO	
Vertical Re-Scaling	yes / no	must equal NO	
Vertical Shift	+/- # of lines	20 lines	
Jitter	yes / no	must equal NO	
Horizontal Cropping	x luma samples	30 pixels	
Vertical Cropping	# of lines	20 lines	

Frame Alignment	yes / no	must equal YES	

ANNEX V
Glossary

To better understand the content of this document it is useful to recall the meaning of some key terms commonly used in the description of formal subjective tests.

Basic test cell: it is the smallest element by which a test is designed; each basic test cell provides the evaluation of one single test condition.

Coding condition: same as HRC.

Contextual effect: the influence in the subjects quality judgement due to the presentation of basic cells in which the quality level of two consecutive ones differs in a consistent way; the reaction of the subject may be driven to improve their response (in a negative or positive direction), compared to the reaction they could have whenever the quality gap is limited.

Formal subjective test: it is the complete experiment during which all the conditions under test are evaluated by means of the all selected original sequences, using non-expert observers.

HRC: Hypothetical Reference Circuit; it represents the process (encoding, transmission and decoding path) to be evaluated.

Instructions to the subjects: a written text that must be read to the subjects before the execution of the training test session, and test session.

Laboratory set-up: the guidelines and the laboratory instrumentation required to properly execute a formal subjective assessment.

Material under test: video material (namely around 10 seconds in length) processed according to the test conditions.

Original sequence: source video material not processed.

Processed sequence: same as “Material under test”.

Randomisation of the test cells: the process by which the order of the basic test cells of one (or more) test session(s) are presented to the subjects; this process tries to avoid biasing of judgement and or possible decrease of attention of the subjects.

Sequence: short video portion (namely around 10 seconds in length).

Stabilisation phase: a number of basic cells (typically five) that represents the whole range of quality of a test session; the stabilisation phase must be present at the start of each test session and must be done using basic cells of the test session it belongs to.

Source material: same as “original sequence”

Subject: is a person that is asked to express his/her subjective opinion of quality during a test session,

Subject training process: is the methodology by which the subjects are instructed about the task they are supposed to do during a formal subjective test; the training process must be carried out **only once** immediately before the beginning of the formal subjective test, and it has not to be repeated to the subjects that carry out more than one test session (or formal test) using the same method and in the same period of time (i.e. with interruptions not longer than one day); the training process foresees the dictation of instructions, the conduction of a training test session and a short time dedicated to question and answers (if any); the experimenter must check the result of the training test session to see if the task has been properly done by each subject.

Subject screening: the procedure by which each subject is checked for visual acuity and colour blindness (according to the ITU-R rec. 500-10); for some specific experiment a contrast sensitivity screening may be also done.

Test condition: it is a combination of a coding conditions (HRCs) applied to a test material; usually a formal subjective test covers all the possible combinations.

Test material: same as “original sequence”

Test session: the period of time during which the subjects are shown (without any interruption) a number of basic test cells; should the length in time required to represent all the basic test cells, be longer than 30 minutes, the basic cells are spread over more than one test session; in any case a test session must not exceed in time the duration of 30 minutes.

Training test session: A small test session, typically from three to five basic cells, in which a representative sample of the artefacts to be assessed are shown to the subjects; it has to be separate from the formal tests.

Unprocessed sequence: same as “original sequence”

Viewing angle: the maximum angle from which the subjects must see the monitor during the test (typically 30° off center of the display)

Viewing distance: the distance measured from the subject’s head and the monitor. It is measured in “H”, where H is the height of the monitor used in the test.